Supplementary Materials for "Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape" by Hie, Cho, DeMeo, Bryson, and Berger

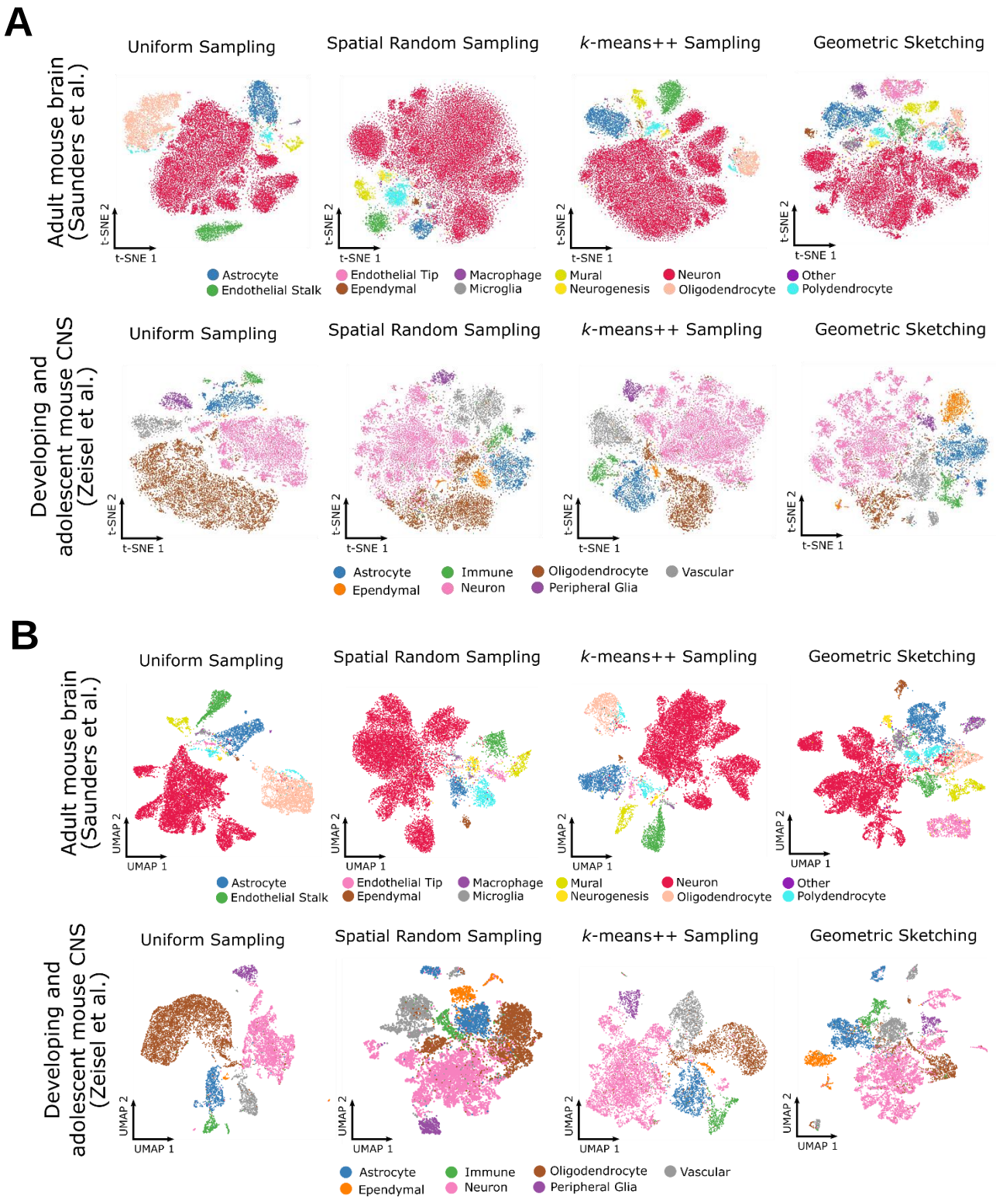

Figure S1: Visualizations of Different Sketches of Large-Scale scRNA-seq Datasets, Related to Figure 3

Visualizations using (**A**) *t*-SNE and (**B**) UMAP of sketches containing 2% of the cells from the adult mouse brain (Saunders et al., 2018) and from the developing and adolescent mouse CNS (Zeisel et al., 2018) using uniform random sampling, SRS, *k*-means++ and geometric sketching. Numbers of cells from each cell type are given in **Table S3-S4**. Note that all data-dependent sampling methods underrepresent oligodendrocytes compared to uniform sampling, which is expected given the low transcriptional heterogeneity among oligodendrocytes as quantified by differential entropy (**Table S3-S4**). While some sketches obtained by *k*-means++ sampling and SRS may appear similar to geometric sketches, they quantifiably preserver fewer rare cell types and have lower sketch quality as measured by the Hausdorff distance, which we show in our experiments.

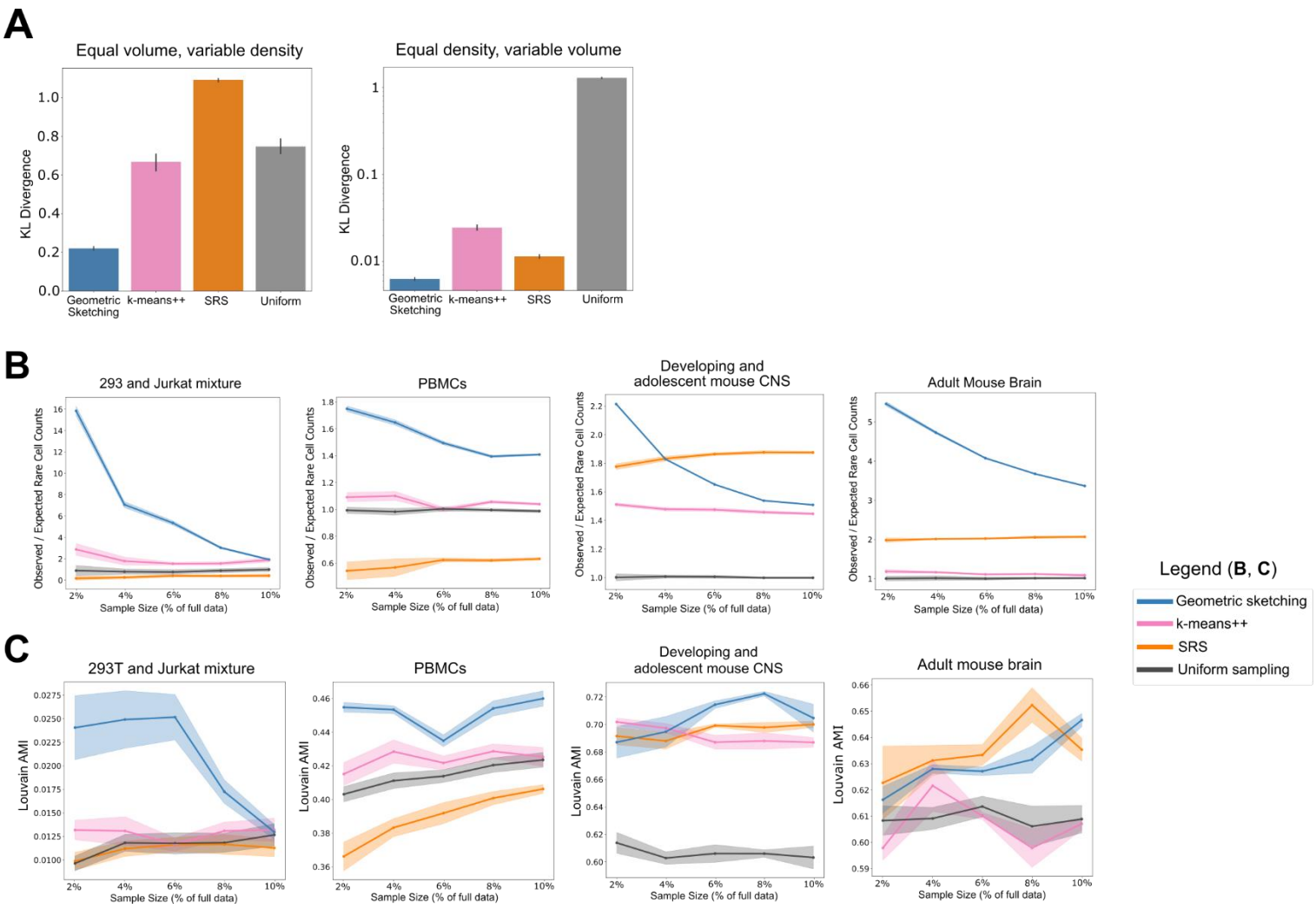

Figure S2: Additional Benchmark Comparisons between Geometric Sketching and Other Sampling Methods, Related to Figure 3

(**A**) Sampling with geometric sketching better reflects differences in cluster volume instead of density**.** Geometric sketching samples from clusters according to the volume of space occupied by each cluster. Bar height indicates means and error bars indicate standard error across 10 random seeds. The *y*-axis indicates the KL divergence of expected cluster representation based on known cluster volumes compared to observed cluster representation in the subsampled data; KL divergences for the equal density, variable volume experiment are plotted on a log scale. Closer to 0 is better (indicates less bias introduced by density). The datasets consist of clusters of equal volume but varying densities or clusters with equal numbers of cells but varying volumes.

(**B**) Rarest cell types are more represented within a geometric sketch. We assessed overrepresentation of cell types within a sketch by computing the ratio of the observed number of cells over the expected number of cells (assuming uniform sampling probability) for each cell type; we then took the geometric mean of the ratios for the rarest half of all cell types within each dataset. Geometric sketching consistently overrepresents rare cell types and does so more than other sampling strategies in almost all cases. Because we set the number of covering boxes equal to the desired sketch size, as the sketch size increases, the overrepresentation ratio with respect to uniform sampling will converge to unity. (**C**) Unbalanced measurement of clustering recapitulation of biological cell types. The same result as in **Figure 3D** but without equal weighting of biological cell types. Louvain clustering was applied to a sketch, transferred to the full dataset, and then measured for agreement with biological cluster labels using adjusted mutual information (**Method Details**). Unsupervised clustering of geometric sketches more consistently recapitulates biological cell types than clustering results obtained by uniform sampling and is comparable to or better than clusters of sketches from *k*-means++ and SRS.

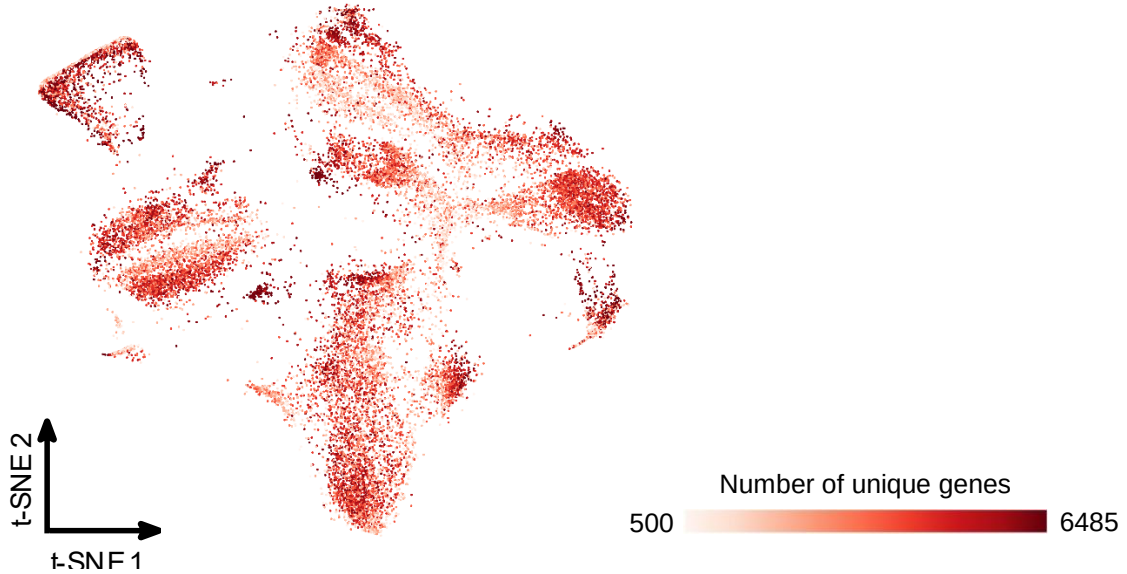

Figure S3: Unique Gene Heatmap of Umbilical Cord Cells, Related to Figure 4

Heatmap of the *t*-SNE embedded geometric sketch visualizing cells from human umbilical cord blood colored by the number of unique genes. Lighter red indicates higher levels of sparsity and darker red indicates lower levels of sparsity. The lowest number of unique genes in the dataset was 500 and the highest was 6485 out of a total of 33,694 genes considered in the study.

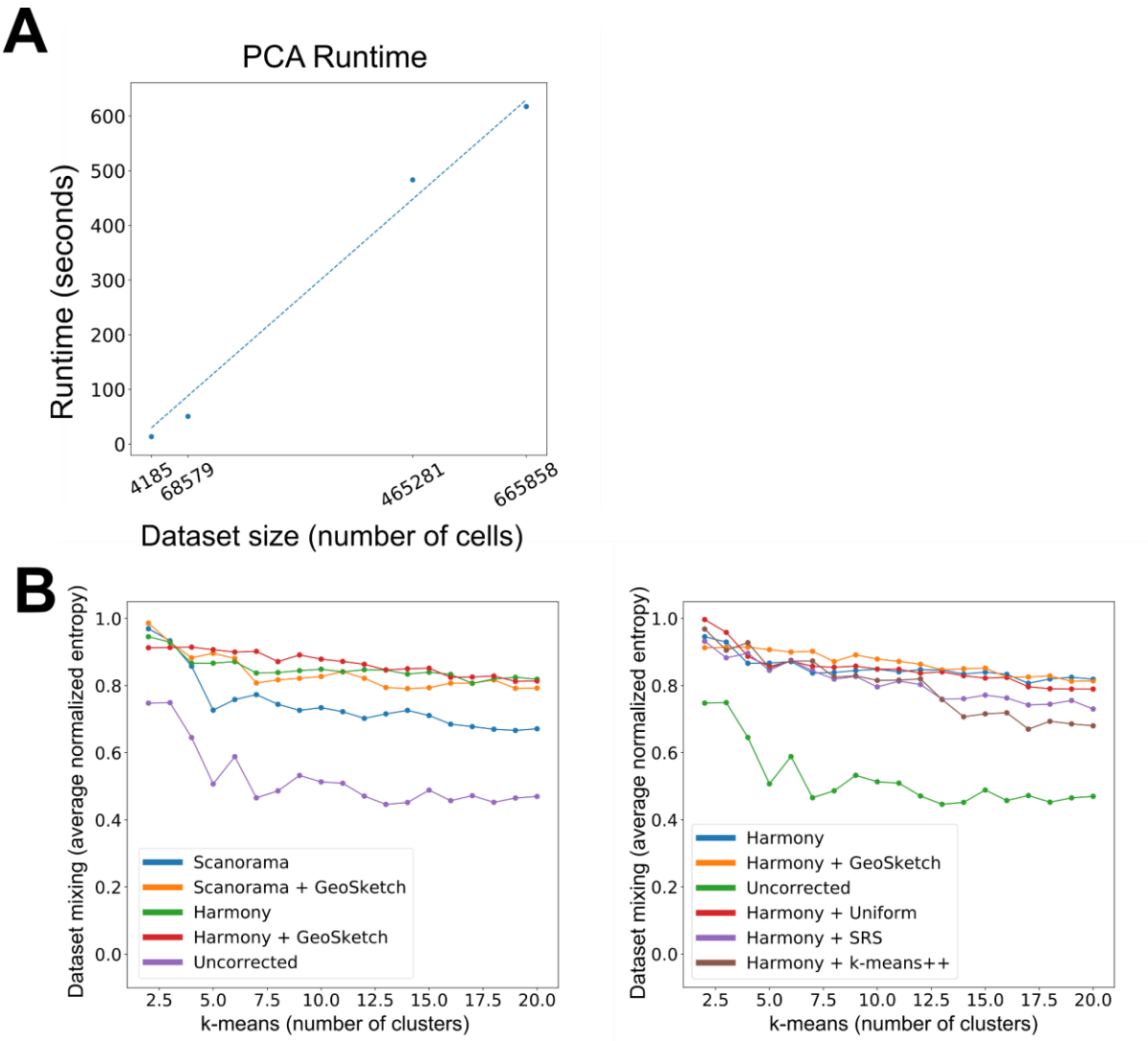

Figure S4: Scalability and Accelerative Capacity of Geometric Sketching, Related to Figure 5

(**A**) PCA runtime versus dataset size. The time required to learn a 100-dimensional representation of a scRNA-seq dataset using a randomized PCA (Halko et al., 2011) scales linearly with the size of the dataset and has reasonable scalability to large-scale scRNA-seq experiments in the future. Each point given in the above plot corresponds to the time taken to compute a 100-dimensional embedding on each of the four main benchmark datasets used in the

study. **(B)** Integration quality of methods with and without geometric sketching-based acceleration.

Closer to 1 indicates more dataset mixing within clusters; see **Method Details** for description of our integration quality metric. Geometric sketching-based acceleration of integration methods yields integrations with comparable or better quality than applying the integration methods to the full dataset. Both geometric sketching and uniform sampling have comparable integration quality, but based on our other results, it is likely that geometric sketching would better align rare cell types in addition to common cell types. Using SRS and *k*-means++ sampling produces worse integration quality.

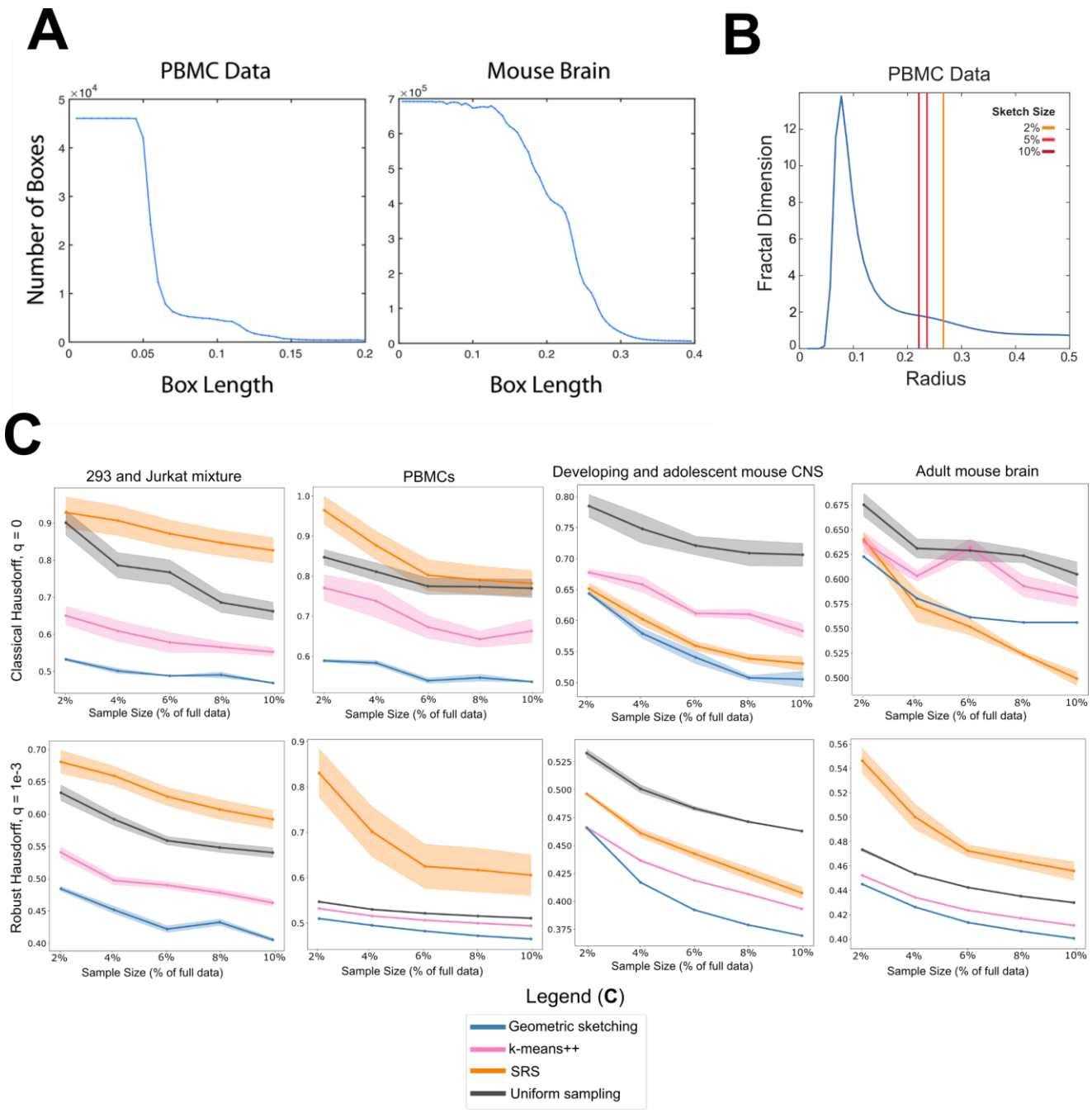

Figure S5: Methodologically Relevant Properties of Geometric Sketching, Related to STAR Methods

(**A**) Near monotonicity of covering boxes with box length. Cardinality of plaid covering near-monotonically decreases with respect to the length parameter. For PBMC and adult mouse brain datasets, we plotted the number of boxes returned by our plaid covering algorithm as a function

of box length provided as input. The overall monotonic relationship allows us to use binary search to find the length at which the plaid cover contains roughly the desired number of boxes. (**B**) Low fractal dimension of single-cell data**.** On the PBMC dataset, we computed the fractal dimension (averaged over the data points) at varying box lengths using the Chebyshev metric, which induces covering spheres that appear as boxes. Letting $N_r(x)$ be the number of data points covered by a sphere of radius r centered at x, we define fractal dimension as $\log\big(N_{r_1}(x)/N_{r_2}(x)\big)/\log(r_1/r_2)$. Plot shows fractal dimension computed over 50 evenly spaced intervals between 0 and 0.5. Various vertical lines denote the radiuses that corresponds to the box size chosen by our geometric sketching algorithm when obtaining sketches containing different percentages of the overall dataset. At the scale at which our geometric sketching operates, PBMC data displays a low fractal dimension of around 2. (**C**) Partial Hausdorff distance at different parameter cutoffs. We measured the partial Hausdorff distance at different values of the parameter q (**Method Details**), including $q =$ 1e-4 (**Figure 3A**), $q =$ 1e-3 and $q =$ 0 (the last corresponding to the classical Hausdorff distance). Geometric sketching outperforms all other sampling methods when measured with a robust, partial Hausdorff distance with positive q. Under the classical Hausdorff distance, geometric sketching also outperforms all other sampling methods in almost all cases except for larger sketches in the adult mouse brain dataset due to a single outlier cell, but the anomalous cell was removed when computing more robust Hausdorff distance measures.

Cell Type	Number of cells	% of total	Differential Entropy
293T	28	0.669056	-461.66
Jurkat	4157	99.33094	-270.88

Table S1, Related to Figure 3

Statistics for 293/Jurkat mixture data; for the differential entropy calculation, see **Method Details**.

Cell Type	Number of cells	% of total	Differential Entropy
CD14+ Monocyte	3817	5.565844	-228.419
CD19+ B	3306	4.820718	-213.47
CD4+/CD25 T	2812	4.100381	-238.942
CD4+/CD45RA+/CD25- Naive T	3126	4.558247	-230.899
CD4+/CD45RO+ Memory	5859	8.543432	-223.313
CD4+ Helper T	11445	16.68878	-222.592
CD56+ NK	14112	20.57773	-232.116
CD8+/CD45RA+ Naive Cytotoxic	21975	32.04334	-232.351
CD8+ Cytotoxic T	1865	2.719491	-219.693
Dendritic	262	0.382041	-281.506

Table S2, Related to Figure 3

Statistics for PBMC data; for the differential entropy calculation, see **Method Details**.

Cell Type	Number of cells	% of total	Differential Entropy	Uniform	*k*-means++	SRS	Geometric sketching
Astrocyte	54444	8.176518	-285.773	1088	1090	389	1277
Endothelial Stalk	39298	5.901859	-271.857	761	782	533	556
Endothelial Tip	3818	0.573396	-277.978	84	107	175	815
Ependymal	2157	0.323943	-282.046	33	68	102	165
Macrophage	1695	0.254559	-290.916	43	31	47	262
Microglia	4614	0.692941	-275.472	86	75	99	397
Mural	12083	1.814651	-270.937	247	297	346	519
Neurogenesis	2372	0.356232	-257.468	47	89	171	151
Neuron	428051	64.28563	-232.534	8655	9746	10821	7975
Oligodendrocyte	104773	15.73504	-342.73	2031	747	53	627
Other (unlabeled)	379	0.056919	-281.542	5	10	24	49
Polydendrocyte	12174	1.828318	-260.35	237	275	557	524

Table S3, Related to Figure 3

The second through fourth columns give the statistics for adult mouse brain data; for the differential entropy calculation, see **Method Details**. The fifth through seventh columns give the number of cells from each cell type in subsamples visualized in **Figure 3B** from Saunders *et al*. (2018).

Cell Type	Number of Cells	% of total	Differential Entropy	Uniform	*k*-means++	SRS	Geometric sketching
Astrocyte	34915	7.504067	-293.16	697	949	905	1194
Ependymal	2777	0.596844	-274.99	49	115	339	614
Immune/Blood	14081	3.026343	-289.20	265	418	350	466
Neuron	147059	31.60649	-243.09	2982	4823	3911	4533
Oligodendrocyte	219220	47.11561	-338.52	4371	1649	2273	1098
Peripheral Glia	16066	3.452967	-328.23	332	322	259	230
Vascular	31163	6.697673	-265.75	609	1029	1268	1170

Table S4, Related to Figure 3

The second through fourth columns give the statistics for developing and adolescent mouse CNS data; for the differential entropy calculation, see **Method Details**. The fifth through seventh columns give the number of cells from each cell type in subsamples visualized in **Figure 3B**.